Probab. Theory Relat. Fields
DOI 10.1007/s00440-008-0180-8

Aggregation via empirical risk minimization

Guillaume Lecué · Shahar Mendelson

Received: 18 December 2007 / Revised: 11 September 2008

Abstract Given a finite set F of estimators, the problem of aggregation is to construct a new estimator whose risk is as close as possible to the risk of the best estimator in F. It was conjectured that empirical minimization performed in the convex hull of F is an optimal aggregation method, but we show that this conjecture is false. Despite that, we prove that empirical minimization in the convex hull of a well chosen, empirically determined subset of F is an optimal aggregation method.

Mathematics Subject Classification (2000) 62G08 · 62C12

1 Introduction

In this article, we first solve a problem concerning aggregation of estimators that was posed by P. Massart. We then construct a new optimal aggregation procedure via empirical risk minimization over the convex hull of an empirically chosen subset.

To formulate the problem we address we need several definitions. Let Ω be a measurable space endowed with a probability measure μ and let F be a finite class of real-valued functions on Ω. Let ν be a probability measure on $\Omega \times \mathbb{R}$ such that μ is its marginal on Ω and put (X, Y) and $\mathcal{D} := (X_i, Y_i)_{i=1}^n$ to be $n + 1$ independent random variables distributed according to ν.

This paper was supported in part by an Australian Research Council Discovery grant DP0559465 and by an Israel Science Foundation grant 666/06.

G. Lecué (✉) · S. Mendelson
Centre for Mathematics and its Applications, The Australian National University,
Canberra, ACT 0200, Australia
e-mail: lecue@latp.univ-mrs.fr

S. Mendelson
e-mail: shahar.mendelson@anu.edu.au

G. Lecué · S. Mendelson
Department of Mathematics, Technion, I.I.T, Haifa 32000, Israel

From a statistical point of view, we want to predict the values of Y at the point X from the observations $\mathcal{D}$. If f is a candidate predictor of Y, the quality of prediction of Y by f is given by the *risk* of f:

$$R(f) = \mathbb{E}\ell(f(X), Y),$$

where $\ell : \mathbb{R}^2 \to \mathbb{R}$ is a nonnegative function, called the *loss function*. If $\hat{f}$ is a random function determined using the data $\mathcal{D}$, the quality of prediction of Y by $\hat{f}$ is the conditional expectation

$$R(\hat{f}) = \mathbb{E}\left(\ell(\hat{f}(X), Y)|\mathcal{D}\right).$$

Throughout this article, we will restrict ourselves to functions f and targets Y that are bounded in L_∞ by b. Also, we will only consider finite classes F of cardinality M.

Given a map (or learning algorithm) A that assigns to a sample $\mathcal{D}$ a function $A_\mathcal{D} \in F$, and for any confidence parameter δ, the *uniform error rate* of A is defined as the function $H(n, M, \delta)$ for which the following holds: for every integer n, every class F of cardinality M and every target Y (all bounded by b), with ν^n-probability at least $1 - \delta$ (i.e. relative to samples of cardinality n),

$$R(A_\mathcal{D}) \leq \min_{f \in F} R(f) + H(n, M, \delta).$$

One can show [10,12,15,17] that, if $\ell(x, y) = (x - y)^2$, then for every random map A there exists a constant c depending only on the map and on δ, such that for every n, $H(n, M, \delta) \geq c/\sqrt{n}$. In fact, the result is even stronger—this lower bound holds for every individual class F that satisfies certain conditions (rather than a lower bound for the "worst" case) and for a wider class of loss functions.

The lower bound on $H(n, M, \delta)$ implies that regardless of the estimation procedure one chooses, it is impossible to obtain error rates that converge to 0 faster than $1/\sqrt{n}$, and that hold uniformly for every F of cardinality M. Thus, to find a procedure that would give faster rates than $1/\sqrt{n}$ one has to consider maps into larger classes than F itself. This leads to the notion of *aggregation* [4,20].

In the aggregation framework, one is given a set F of M functions (usually selected in a preliminary stage out of a larger class as potentially good estimators of Y). The problem of aggregation is to construct a procedure that mimics the best element in F, without the restriction that A has to select a function in F itself. Having this in mind, one can define the *optimal rate of aggregation* [22], in a similar way to the notion of the minimax rate of convergence for the estimation problem. This is the smallest price that one has to pay to mimic, in expectation, the best element in a function class F of cardinality M from n observations. Here, we focus on results that hold with high probability and consider the following definition of optimality.

Definition 1.1 A function $\psi(n, M)$ is an optimal rate of aggregation with confidence $0 < \delta < 1/2$ and a procedure $\tilde{f}$ is an optimal aggregation procedure with confidence δ if there exists a constant $c_1(\delta)$ for which the following hold:

- For any integers n and M, any set F of cardinality M and any target Y (all bounded by b), with ν^n-probability at least $1-\delta$, $\tilde{f}$ satisfies

$$R(\tilde{f}) \leq \min_{f\in F} R(f) + c_1(\delta)\psi(n, M),$$

  where $R(\tilde{f})$ is the conditional expectation $\mathbb{E}(\ell(\tilde{f}(X), Y)|\mathcal{D})$.
- There exists an absolute constant $c_2 > 0$ such that the following holds. For any integers M and n, and any procedure $\bar{f}$, there exist a set F of cardinality M and a target Y (all bounded by b) such that, with ν^n-probability at least $1/2$,

$$R(\bar{f}) \geq \min_{f\in F} R(f) + c_2\psi(n, M).$$

One can show [4,10,22] that if the loss satisfies a slightly stronger property than convexity, then the optimal rate of aggregation (in the sense of the definition in [22]) is

$$\frac{\log M}{n}. \tag{1.1}$$

This is significantly better than the rate of $\sqrt{(\log M)/n}$, which is, in general, the best rate that one can obtain when A is restricted to F itself.

Note that standard lower bounds on aggregation rates do not hold with large probability; they are given in expectation, following the definition of [23]. Nevertheless, by using the same classical tools as in Chapter 2 of [23], it is easy to prove that the second point of Definition 1.1 is fulfilled with the aggregation rate of $\psi(n, M) = (\log M)/n$.

A natural procedure that is very useful in prediction is *empirical risk minimization*, which assigns to each sample $\mathcal{D}$ a function that minimizes the *empirical risk*

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)$$

on a given set.

One can show that for the type of classes we have in mind—finite, of cardinality M—empirical minimization in F gives the optimal rate among all the maps A restricted to taking values in F. Since there are optimal aggregation procedures that are convex combinations of elements in F [4,10], it is natural to believe that empirical risk minimization performed in the convex hull of F rather than in F itself would achieve the optimal rate of aggregation (1.1). This was the question asked by P. Massart.

Question 1.2 *Is empirical minimization performed on* $\mathrm{conv}(F)$ *an optimal aggregation procedure?*

In addition to its theoretical value, Question 1.2 is highly motivated by practical considerations. Indeed, convex aggregation methods typically depend on some tuning parameter (such as the temperature in the case of a Gibbs prior and a Bayesian aggregation procedure [4,10]. Of course, one is led to choose the tuning parameter that

minimizes the empirical risk, and this choice turns out to be a rather efficient one, as shown by some empirical studies in [7]. Nevertheless, there is no known theoretical result on the choice of the tuning parameter.

One is tempted to expect a positive answer to Question 1.2 because the approximation error $\inf_{f\in \mathrm{conv}(F)} R(f)$ is likely to be significantly smaller than $\min_{f\in F} R(f)$ - to a degree that outweighs the increased statistical error caused by performing empirical minimization in the much larger set $\mathrm{conv}(F)$. However, our first result is that contrary to this intuition, the answer to Question 1.2 is negative in a very strong way.

Theorem A *There exist absolute constants c_1, c_2 and c_3 for which the following holds. For every integer n there is a set F_n of functions of cardinality $M = c_1\sqrt{n}$ and a target Y (all bounded by 1), such that with ν^n-probability of at least $1 - \exp(-c_2\sqrt{n})$,*

$$R(\hat{f}) \geq \min_{f\in F_n} R(f) + \frac{c_3}{\sqrt{n}},$$

where $\hat{f}$ is the empirical minimizer in $\mathrm{conv}(F)$ and R is measured relative to the squared loss $\ell(x, y) = (x - y)^2$.

In other words, empirical minimization performed in $\mathrm{conv}(F)$ does not even come close to the optimal aggregation rate. In fact, it is not far from the trivial rate that one can achieve by performing empirical minimization in F itself - which is $c(\delta)\sqrt{(\log M)/n}$.

Nevertheless, understanding why empirical minimization does not perform well on F and on $\mathrm{conv}(F)$ as an aggregation method does lead to an improved procedure. Our second result shows that empirical minimization on an appropriate, data dependent subset of $\mathrm{conv}(F)$ achieves the optimal rate of aggregation in (1.1). To formulate our result, denote for every sample $\mathcal{D} = (X_i, Y_i)_{i=1}^{2n}$, the subsamples $D_1 = (X_i, Y_i)_{i=1}^{n}$ and $D_2 = (X_i, Y_i)_{i=n+1}^{2n}$. For every $x > 0$, let $\alpha = ((x + \log M)/n)^{1/2}$, and for every sample $\mathcal{D} = (X_i, Y_i)_{i=1}^{2n}$, set

$$\hat{F}_1 = \left\{ f \in F : R_n(f) \leq R_n(\hat{f}) + C_1 \max\left\{ \alpha \|\hat{f} - f\|_{L_2^n}, \alpha^2 \right\} \right\}, \tag{1.2}$$

where $\hat{f}$ is an empirical minimizer in F with respect to D_1, L_2^n is the L_2 space with respect to the random empirical measure $n^{-1}\sum_{i=1}^n \delta_{X_i}$ and $C_1 > 0$ is a constant depending only on ℓ and b.

Theorem B *Under mild assumptions on the loss ℓ, there exists a constant c_1 depending only on b and ℓ for which the following holds. Let F be a class of functions bounded by b and of cardinality M and assume that Y is bounded by b.*

For any $x > 0$, if $\tilde{f}$ is the empirical minimizer in the convex hull of $\hat{F}_1$ with respect to D_2 then, with ν^{2n}-probability at least $1 - 2\exp(-x)$,

$$R(\tilde{f}) \leq \min_{f\in F} R(f) + c_1(x + 1)\frac{\log M}{n}.$$

It is important to mention that the boundedness assumption on Y is crucial to the analysis we present here. Nevertheless, it is possible to obtain similar results under milder assumptions on Y and F, using the methods developed in [9,18,19] for handling empirical processes indexed by powers of unbounded function classes that satisfy suitable tail assumptions. Since the analysis of the unbounded case is technically much harder and would shift the emphasis of this article away from the main ideas we wish to present, we will only consider the uniformly bounded case. Also, there are some known results in the unbounded case, where the assumption that F and Y are uniformly bounded has been avoided by other methods (for example, convex optimization). All the known procedures in that direction rely on exponential aggregating schemes that depend on an unknown tuning parameter and some convexity assumptions on the loss function, and are very different from what we do. Moreover, these estimates hold only in expectation rather than with exponential probability as we have here. Indeed, in [4,10], optimal inequalities of the same flavor as Theorem B have been obtained (in expectation), as well as in [2], where the results are optimal up to a logarithm factor, while in [5], PAC-Bayesian bounds [4] have been obtained with respect to the empirical risk $\bar{R}_n(f) = (1/n)\sum_{i=1}^n f^2(X_i)$. One should note that unlike the aggregation procedures that are based on exponential weights, our procedure enjoys some "sparsity" properties, in the sense that elements of the dictionary that are not relevant do not appear in the final aggregate. This property is a consequence of our pre-selection step which, we hope, might be used to solve some practical and theoretical problems relating to sparsity.

The geometric motivation behind the proof of Theorem B will be explained in the following section and the proof of the theorem will appear in Sect. 4. We will present the proof of Theorem A in Sect. 5.

Finally, a word about notation. Throughout, we denote absolute constants or constants that depend on other parameters by c_1, c_2, etc., (and, of course, we will specify when a constant is absolute and when it depends on other parameters). The values of constants may change from line to line. Constants whose value will remain fixed are denoted by C_1, C_2, etc. Given a sample $(Z_i)_{i=1}^n$, we set $P_n = n^{-1}\sum_{i=1}^n \delta_{Z_i}$, the random empirical measure supported on $\{Z_1, \ldots, Z_n\}$. For any function f let $(P_n - P)(f) = n^{-1}\sum_{i=1}^n f(Z_i) - \mathbb{E}f(Z)$ and for a class of functions F, $\|P_n - P\|_F = \sup_{f\in F} |(P_n - P)(f)|$.

2 The role of convexity in aggregation

In this section, our goal is to explain the geometric idea behind the proofs of Theorems A and B. To that end, we will restrict ourselves to the case of the squared loss $\ell(x, y) = (x - y)^2$ and a noiseless target function $T : \Omega \to \mathbb{R}$.

Set $f^F = \operatorname{argmin}_{f\in F} \mathbb{E}\ell(f, T)$ and observe that f^F minimizes the $L_2(\mu)$ distance between T and F. Recall that our aim is to construct some $\tilde{f}$, such that with probability at least $1 - \delta$,

$$\|\tilde{f} - T\|_{L_2(\mu)}^2 \leq \|f^F - T\|_{L_2(\mu)}^2 + c(\delta)\Phi(n, M),$$

where n is the sample size, the cardinality of F is $|F| = M$ and $\Phi(n, M)$ is as small as possible—hopefully, of the order of $n^{-1} \log M$.

The motivation to select $\tilde{f}$ from $\mathcal{C} = \mathrm{conv}(F)$ is natural, since one can expect that $\min_{h \in \mathcal{C}} \|h - T\|_{L_2(\mu)} = \|f^{\mathcal{C}} - T\|_{L_2(\mu)}$ is much smaller than $\|f^F - T\|_{L_2(\mu)}$. Moreover, it is reasonable to think that empirical minimization performed in $\mathcal{C}$ has a relatively fast error rate, which we denote by $c_1(\delta)\Psi(n, M)$. Therefore, if $\tilde{f}$ is the empirical minimizer in $\mathcal{C}$ then

$$\begin{aligned}
&\|\tilde{f} - T\|^2_{L_2(\mu)} \\
&\quad \le \|f^{\mathcal{C}} - T\|^2_{L_2(\mu)} + c_1(\delta)\Psi(n, M) \\
&\quad \le \|f^F - T\|^2_{L_2(\mu)} + c_1(\delta)\Psi(n, M) - \Big(\|f^F - T\|^2_{L_2(\mu)} - \|f^{\mathcal{C}} - T\|^2_{L_2(\mu)}\Big),
\end{aligned}$$

and the hope is that the gain in the approximation error

$$\Big(\|f^F - T\|^2_{L_2(\mu)} - \|f^{\mathcal{C}} - T\|^2_{L_2(\mu)}\Big)$$

is far more significant than $\Psi(n, M)$, leading to a very fast aggregation rate.

Although this approach is tempting, it has serious flaws. First of all, it turns out that the statistical error of empirical minimization in a convex hull of M well chosen functions may be as bad as $1/\sqrt{n}$ (for $M \sim \sqrt{n}$, see Theorem 5.5). Second, it is possible to construct such a class and a target for which $\|f^F - T\|_{L_2(\mu)} = \|f^{\mathcal{C}} - T\|_{L_2(\mu)}$, and thus, there is no gain in the approximation error by passing to the convex hull.

The class we shall construct will be $\{0, \pm\phi_1, \ldots, \pm\phi_M\}$ where $(\phi_i)_{i=1}^M$ is a specific orthonormal family on $[0, 1]$ and the target Y is ϕ_{M+1}, implying that $f^F = f^{\mathcal{C}} = 0$. For this choice of F and Y one can show that $\Psi(n, c_1\sqrt{n}) \ge c_2/\sqrt{n}$ for suitable absolute constants c_1 and c_2.

Fortunately, not all is lost as far as using empirical minimization in a convex hull, but one has to be more careful in selecting the set in which it is performed. The key point is to identify situations in which there is a significant gain in the approximation error by passing to the convex hull.

Assume that there are at least two functions in F that almost minimize the loss R in F (which, in the square loss case, is the same as almost minimizing the L_2 distance between T and F), and that these two functions are relatively "far away" from each other in L_2. By the parallelogram equality (or by a uniform convexity argument for a more general loss function), if f_1 and f_2 are "almost minimizers" then

$$\begin{aligned}
\left\| \frac{f_1 + f_2}{2} - T \right\|^2_{L_2(\mu)} &\le \frac{1}{2}\|f_1 - T\|^2_{L_2(\mu)} + \frac{1}{2}\|f_2 - T\|^2_{L_2(\mu)} - \frac{1}{4}\|f_1 - f_2\|^2_{L_2(\mu)} \\
&\approx \|f^F - T\|^2_{L_2(\mu)} - \frac{1}{4}\|f_1 - f_2\|^2_{L_2(\mu)}.
\end{aligned}$$

Thus, if F_1 is the set of all the almost minimizers in F of the distance to T and the diameter of F_1 is large (to be precise, larger than $c/\sqrt{n}$), the approximation error in the convex hull of F_1 is significantly smaller than in F. On the other hand, one can show that if the diameter of F_1 is smaller that $c/\sqrt{n}$, the empirical minimization algorithm in $\mathrm{conv}(F_1)$ has a very fast error rate (because one has a very strong control on the variances of the various loss functions associated with this set). Therefore, in both cases-but for two completely different reasons—if $\tilde{f}$ is the empirical minimizer performed in the convex hull of F_1 then $\|\tilde{f}-T\|^2_{L_2(\mu)} \leq \|f^F-T\|^2_{L_2(\mu)}+c(\delta)(\log M)/n$, with probability greater than $1-\delta$.

Naturally, using F_1 is not realistic because it is impossible to identify the set of almost true minimizers of the risk in F using the given data. However, it turns out that one can replace F_1 with a set that can be determined empirically and has similar properties to F_1. The set defined in (1.2) satisfies that if its $L_2(\mu)$ diameter is larger than $c/\sqrt{n}$ then the gain in the approximation error in its convex hull is dramatic (compared with the one in F), while if its diameter is smaller than $c/\sqrt{n}$ then empirical minimization performed in its convex hull yields a very fast error rate.

3 Preliminaries from empirical processes theory

Here, we will present some of the results we need for our analysis, the first of which is Talagrand's concentration inequality for empirical processes indexed by a class of uniformly bounded functions.

Theorem 3.1 [13] *Let F be a class of functions defined on (Ω,μ) such that for every $f\in F$, $\|f\|_\infty \leq b$ and $\mathbb{E}f=0$. Let $X_1,\ldots,X_n$ be independent random variables distributed according to μ and set $\sigma^2 = n\sup_{f\in F}\mathbb{E}f^2$. Define*

$$Z=\sup_{f\in F}\sum_{i=1}^n f(X_i) \quad \text{and} \quad \bar{Z}=\sup_{f\in F}\left|\sum_{i=1}^n f(X_i)\right|.$$

Then, for every $x>0$ and every $\rho>0$,

$$Pr\left(\left\{Z\geq(1+\rho)\mathbb{E}Z+\sigma\sqrt{Kx}+K(1+\rho^{-1})bx\right\}\right)\leq e^{-x},$$

$$Pr\left(\left\{Z\leq(1-\rho)\mathbb{E}Z-\sigma\sqrt{Kx}-K(1+\rho^{-1})bx\right\}\right)\leq e^{-x},$$

and the same inequalities hold for $\bar{Z}$.

In our discussion, we will be interested in empirical processes indexed by a finite class of functions F and in excess loss classes associated with F or with its convex hull, which is denoted by $\mathcal{C}$.

Given a class $\mathcal{G}$, the excess loss function associated with $\mathcal{G}$ and a function h is

$$\mathcal{L}_{\mathcal{G}}(h)(X,Y) = \ell(h(X),Y) - \ell(h^{\mathcal{G}}(X),Y),$$

where $h^{\mathcal{G}}$ minimizes $h \longmapsto \mathbb{E}\ell(h(X),Y)$ in $\mathcal{G}$. Let $\mathcal{L}_{\mathcal{G}}(F) = \{\mathcal{L}_{\mathcal{G}}(f) : f \in F\}$ be the excess loss class relative to $\mathcal{G}$ with a base class F.

In cases where the class $\mathcal{G}$ is clear and $\mathcal{G} = F$, we denote the excess loss class by $\mathcal{L}$ and the excess loss function of h by $\mathcal{L}_h$.

The following lemma is rather standard and we present its proof for the sake of completeness.

Lemma 3.2 *There exist absolute constants c_1, c_2 and c_3 for which the following holds. If F is a finite class of functions bounded by b and*

$$d(F) = \operatorname{diam}(F, L_2(\mu)) \quad \text{and} \quad \sigma^2(F) = \sup_{f\in F} \mathbb{E}f^2,$$

then

$$\mathbb{E}\sup_{f\in F} \frac{1}{n}\sum_{i=1}^n f^2(X_i) \le c_1 \max\left\{\sigma^2(F), b^2\frac{\log|F|}{n}\right\}. \tag{3.1}$$

Also, assume that the target Y is also bounded by b and that the loss ℓ is a Lipschitz function on $[-b,b]^2$ with a constant $\|\ell\|_{\text{lip}}$. If $\mathcal{C} = \operatorname{conv}(F)$ and $\mathcal{H} = \mathcal{L}_{\mathcal{C}}(\mathcal{C})$ then

$$\mathbb{E}\|P_n - P\|_{\mathcal{H}} \le c_3\|\ell\|_{\text{lip}} \max\left\{d(F)\cdot\sqrt{\frac{\log|F|}{n}}, b\frac{\log|F|}{n}\right\}. \tag{3.2}$$

Proof By the Giné–Zinn symmetrization argument [8], the fact that a Bernoulli process is subgaussian with respect to the Euclidean metric and an entropy integral argument (see, for example, [6,14,21,24]) it is evident that for any class F,

$$\mathbb{E}\|P_n - P\|_F \le \frac{2}{n}\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{f\in F}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right| \le \frac{c_1}{\sqrt{n}}\mathbb{E}_X\int_0^r \sqrt{\log N(\varepsilon, F, L_2^n)}d\varepsilon, \tag{3.3}$$

where L_2^n is the L_2 structure with respect to the random empirical measure $n^{-1}\sum_{i=1}^n \delta_{X_i}$ and

$$r^2 = \sup_{f\in F}\frac{1}{n}\sum_{i=1}^n f^2(X_i).$$

Set $F^2 = \{f^2 : f \in F\}$ and notice that by a symmetrization argument and the contraction principle for Bernoulli processes [14, Chapter 4],

$$\begin{aligned}\mathbb{E}r^2 &\leq \left(\mathbb{E}\|P_n - P\|_{F^2}\right) + \sup_{f\in F}\mathbb{E}f^2 \\ &\leq c_2\|F\|_\infty \mathbb{E}\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| + \sigma^2(F),\end{aligned}$$

where $\|F\|_\infty = \sup_{f\in F}\|f\|_\infty$.

Now, if F is a finite class of bounded functions then by setting

$$E = \mathbb{E}\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|$$

and applying (3.3), it is evident that

$$E \leq \frac{c_1}{\sqrt{n}}\sqrt{\log|F|}(\mathbb{E}_X r) \leq \frac{c_3}{\sqrt{n}}\sqrt{\log|F|}\left(\|F\|_\infty E + \sigma^2(F)\right)^{1/2}.$$

Thus,

$$E \leq c_4 \max\left\{\sigma(F)\sqrt{\frac{\log|F|}{n}}, \|F\|_\infty\frac{\log|F|}{n}\right\},$$

and it follows that

$$\mathbb{E}\sup_{f\in F}\frac{1}{n}\sum_{i=1}^n f^2(X_i) \leq c_5\max\left\{\sigma^2(F), \|F\|_\infty^2\frac{\log|F|}{n}\right\}$$

as claimed.

Turning to the second part of the Lemma, let $\mathcal{C} = \mathrm{conv}(F)$, set $\mathcal{H} = \mathcal{L}_{\mathcal{C}}(\mathcal{C})$ and for each $u \in \mathcal{C}$, put $\mathcal{L}_{\mathcal{C}}(u) = \mathcal{L}_u$. Also, denote by $f^{\mathcal{C}}$ the minimizer of $h \longmapsto \mathbb{E}\ell(h(X), Y)$ in $\mathcal{C}$.

Recall that [14, Chapter 4] there exists an absolute constant c_6 such that for every $T \subset \mathbb{R}^n$,

$$\mathbb{E}\sup_{t\in T}\left|\sum_{i=1}^n \varepsilon_i t_i\right| \leq c_6\mathbb{E}\sup_{t\in T}\left|\sum_{i=1}^n g_i t_i\right|,$$

where $(g_i)_{i=1}^n$ are independent, standard Gaussian variables. Hence, for every $(X_i, Y_i)_{i=1}^n$,

$$\mathbb{E}_\varepsilon\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n \varepsilon_i h(X_i, Y_i)\right| \leq c_6\mathbb{E}_g\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n g_i h(X_i, Y_i)\right|.$$

Consider the Gaussian process $v \to Z_v \equiv \sum_{i=1}^n g_i \mathcal{L}_v(X_i, Y_i)$ indexed by $\mathcal{C}$. For every $v, u \in \mathcal{C}$,

$$\begin{aligned}\mathbb{E}|Z_u - Z_v|^2 &= \sum_{i=1}^n (\mathcal{L}_u(X_i, Y_i) - \mathcal{L}_v(X_i, Y_i))^2 \\ &\le \|\ell\|_{\rm lip}^2 \sum_{i=1}^n \left(\left(u - f^{\mathcal{C}}\right) - \left(v - f^{\mathcal{C}}\right)\right)^2 (X_i) \\ &= \|\ell\|_{\rm lip}^2 \mathbb{E}|Z'_u - Z'_v|^2,\end{aligned}$$

where $Z'_u \equiv \sum_{i=1}^n g_i (u - f^{\mathcal{C}})(X_i)$. Therefore, by Slepian's Lemma [6,14], for every $(X_i, Y_i)_{i=1}^n$,

$$\mathbb{E}_g \sup_{h \in \mathcal{H}} \left| \sum_{i=1}^n g_i h(X_i, Y_i) \right| \le \|\ell\|_{\rm lip} \mathbb{E}_g \sup_{v \in {\rm conv}(F)} \left| Z'_v \right| = \|\ell\|_{\rm lip} \mathbb{E}_g \sup_{f \in F} \left| Z'_f \right|.$$

Hence by (3.3) and (3.1) for the class $\{(f - f^{\mathcal{C}}) : f \in F\}$,

$$\begin{aligned}\mathbb{E}\|P_n - P\|_{\mathcal{H}} &\le c_7 \|\ell\|_{\rm lip} \sqrt{\frac{\log|F|}{n}} \left(\mathbb{E}_X \sup_{f \in F} \|f - f^{\mathcal{C}}\|_{L_2^n}^2 \right)^{1/2} \\ &\le c_8 \|\ell\|_{\rm lip} \sqrt{\frac{\log|F|}{n}} \max\left\{ \sup_{f \in F} \|f - f^{\mathcal{C}}\|_{L_2(\mu)}, b\sqrt{\frac{\log|F|}{n}} \right\} \\ &\le c_9 \|\ell\|_{\rm lip} \sqrt{\frac{\log|F|}{n}} \max\left\{ d(F), b\sqrt{\frac{\log|F|}{n}} \right\},\end{aligned}$$

since by convexity, $\sup_{f \in F} \|f - f^{\mathcal{C}}\|_{L_2(\mu)} \le d(F)$. □

Lemma 3.2 combined with Theorem 3.1 leads to the following corollary.

Corollary 3.3 *There exists an absolute constant c for which the following holds. Let F be a finite class of functions bounded by b. For every $x > 0$ and any integer n, let $\alpha = \sqrt{(x + \log|F|)/n}$ and set $d(F) = {\rm diam}(F, L_2(\mu))$. If $\mathcal{C}$ is the convex hull of F and ℓ and $\mathcal{L}$ are as above, then with probability at least $1 - \exp(-x)$, for every $v \in \mathcal{C}$,*

$$\left| \frac{1}{n} \sum_{i=1}^n \mathcal{L}_{\mathcal{C}}(v)(X_i, Y_i) - \mathbb{E}\mathcal{L}_{\mathcal{C}}(v)(X, Y) \right| \le c\|\ell\|_{\rm lip} \max\left\{ \alpha \cdot d(F), b\alpha^2 \right\}.$$

Sketch of the proof. We apply Theorem 3.1 to the process

$$Z = \sup_{v \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^n \mathcal{L}_{\mathcal{C}}(v)(X_i, Y_i) - \mathbb{E}\mathcal{L}_{\mathcal{C}}(v)(X, Y) \right|,$$

and (3.2) provides an upper bound for the expectation $\mathbb{E}Z$. The proof now follows from a simple computation. □

The final result we need follows immediately from Bernstein's inequality [24] combined with a union bound over the finite set F. Because the proof is standard we will not present it here.

Lemma 3.4 *There exists an absolute constant c for which the following holds. Consider F and α as above and for every $f \in F$, let $\mathcal{L}_f(X,Y) = \ell(f(X),Y) - \ell(f^F(X),Y)$, where f^F minimizes $h \longmapsto \mathbb{E}\ell(h(X),Y)$ in F. Then, with probability at least $1-2\exp(-x)$, for every $f \in F$,*

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_f(X_i,Y_i) - \mathbb{E}\mathcal{L}_f(X,Y)\right| \leq c\|\ell\|_{\mathrm{lip}} \max\left\{d_f\alpha, b\alpha^2\right\},$$

where $d_f = \|f - f^F\|_{L_2(\mu)}$. Also, with probability at least $1-2\exp(-x)$, for every $f,g \in F$,

$$\left|\|f-g\|_{L_2^n}^2 - \|f-g\|_{L_2(\mu)}^2\right| \leq c\max\left\{\|f-g\|_{L_2(\mu)}b\alpha, b^2\alpha^2\right\}.$$

4 The optimal aggregation procedure

Throughout this section, we will assume that F is a class of M functions bounded by b. We will also need certain assumptions on the loss ℓ and to that end, recall the following definition, which originated from the notion of uniform convexity of normed spaces.

Definition 4.1 [1] Let $\phi : \mathbb{R} \to \mathbb{R}$ and set $\Psi_\phi : L_2(\Omega \times \mathbb{R}) \to \mathbb{R}$ by

$$\Psi_\phi(f) = \mathbb{E}\phi(f(X,Y)).$$

We say that Ψ_ϕ is uniformly convex with respect to the $L_2(\nu)$ norm if the function δ_ϕ, defined by

$$\delta_\phi(\epsilon) = \inf_{\substack{f,g\in L_2(\nu)\\ \|f-g\|_2\geq\varepsilon}}\left\{\frac{\Psi_\phi(f)+\Psi_\phi(g)}{2} - \Psi_\phi\left(\frac{f+g}{2}\right)\right\} \tag{4.1}$$

is positive for every $\varepsilon > 0$. The function δ_ϕ is called the modulus of convexity of Ψ_ϕ.

For instance, if $\phi(x) = x^2$ then for every $f \in L_2(\nu)$, $\Psi_\phi(f) = \|f\|_{L_2(\nu)}^2$. Thus, using the parallelogram equality, for every $\varepsilon > 0$, $\delta_\phi(\varepsilon) = \varepsilon^2/4$. Note that the assumption that $\delta_\phi(\varepsilon) \geq c_\phi\varepsilon^2$ for every $\varepsilon > 0$ is a quantitative way of ensuring that the functional $\Psi_\phi : L_2(\nu) \longmapsto \mathbb{R}$ enjoys some convexity properties that are close to the parallelogram equality satisfied by the quadratic function risk $f \longmapsto \|f\|_{L_2(\nu)}^2$.

Assumption 4.1 Assume that ℓ is a Lipschitz function on $[-b,b]^2$ with a constant $\|\ell\|_{\rm lip}$. Assume further that there exists a convex function $\phi:\mathbb{R}\to\mathbb{R}^+$ such that for any $f,g\in L_2(\nu)$,

$$\mathbb{E}_\nu \ell(f,g)=\mathbb{E}_\nu\phi(f-g)$$

and that the modulus of convexity δ_ϕ of $\Psi_\phi: f\to \mathbb{E}_\nu\phi(f)$ satisfies $\delta_\phi(\varepsilon)\geq c_\phi\varepsilon^2$ for every $\varepsilon>0$.

In particular, if $\ell(x,y)=(x-y)^2$ then $\delta_\phi(\varepsilon)\geq \varepsilon^2/4$.

We will denote $\ell_f=\ell(f(X),Y)$, $R(f)=\mathbb{E}\ell_f$ and if $\hat{f}$ is a function of the sample $\mathcal{D}$ then $R(\hat{f})=\mathbb{E}\left(\ell_{\hat{f}}|\mathcal{D}\right)$. Also, for $h:(\Omega\times\mathbb{R},\nu)\to\mathbb{R}$, set $P_nh=n^{-1}\sum_{i=1}^n h(X_i,Y_i)$, where $(X_i,Y_i)_{i=1}^n$ are independent, selected according to ν.

Finally, recall that $\alpha=((x+\log M)/n)^{1/2}$ where x is the desired confidence.

The procedure we have in mind is as follows. We consider a sample $\mathcal{D}=(X_i,Y_i)_{i=1}^{2n}$ and split it into two sub-samples, $D_1=(X_i,Y_i)_{i=1}^n$ and $D_2=(X_i,Y_i)_{i=n+1}^{2n}$. We use D_1 to define a random subset of F:

$$\hat{F}_1=\left\{f\in F: R_n(f)\leq R_n(\hat{f})+C_1\max\left\{\alpha\|\hat{f}-f\|_{L_2^n},\alpha^2\right\}\right\}, \tag{4.2}$$

where C_1 is a constant to be named later and that depends only on $\|\ell\|_{\rm lip}$ and b, $R_n(f)=n^{-1}\sum_{i=1}^n\ell(f(X_i),Y_i)$, $\hat{f}$ is a minimizer of the empirical risk $R_n(\cdot)$ in F and L_2^n is the L_2 space endowed by the random empirical measure $n^{-1}\sum_{i=1}^n\delta_{X_i}$.

To make the exposition of our results easier to follow, we avoided presenting the computation of explicit values of constants. Our analysis showed that one can take $C_1=4\|\ell\|_{\rm lip}(1+9b)$—which, of course, is not likely to be the optimal choice of C_1.

The second step in the algorithm is performed using the second part of the sample $\mathcal{D}$. The algorithm produces the empirical minimizer (relative to D_2) of ℓ in the convex hull of $\hat{F}_1$. Let us denote this minimizer by $\tilde{f}$, that is

$$\tilde{f}=\operatorname{argmin}_{h\in{\rm conv}(\hat{F}_1)}\frac{1}{n}\sum_{i=n+1}^{2n}\ell(h(X_i),Y_i).$$

Note that considering only the "significant" part of a given class (like we do by using the subset $\hat{F}_1\subset F$) is an idea that already appeared, for example, in [16]. In that article, the authors used this idea to construct a very sharp data-dependent penalty function which outperforms most of the well known data-dependent penalties like local Rademacher penalties (see [11] and reference therein) that are usually computed over the entire class. However, this type of "random subset" is different from the one we introduce here. Usually, the random subset consists of functions for which the empirical loss is smaller than the sum of the loss of the empirical minimizer and a sample-dependent complexity term; this complexity term does not depend on each $f\in F$, but rather, on the entire set. Here, in place of the complexity term we use a metric condition: that the empirical L_2 distance between a function and the empirical minimizer.

The main result of this section was formulated in Theorem B.

Theorem 4.2 *For every b and $\|\ell\|_{\mathrm{lip}}$ there exists a constant c_1, depending only on b and $\|\ell\|_{\mathrm{lip}}$, for which the following holds. For any $x > 0$, every class F of M functions, any target Y (all bounded by b) and any loss ℓ satisfying Assumption 4.1, the empirical minimizer $\tilde{f}$ over the convex hull of $\hat{F}_1$ satisfies, with ν^{2n}-probability at least $1 - 2\exp(-x)$,*

$$R(\tilde{f}) = \mathbb{E}\left(\ell_{\tilde{f}}(X, Y) | (X_i, Y_i)_{i=1}^{2n}\right) \leq \min_{f \in F} R(f) + c_1(1 + x)\frac{\log M}{n},$$

Remark 4.3 Note that the definition of the set $\hat{F}_1$, and thus the algorithm, depends on the confidence x one is interested in through the factor α. Thus $\tilde{f}$ also depends on the confidence.

Theorem 4.2 and the fact that $(\log M)/n$ is the best rate one can hope for proves our optimality claim. One can take $c_1(\delta) = c_1(1 + \log(2/\delta))$ for the constant introduced in Definition 1.1.

The idea of the proof is based on constructing a set of "almost minimizers" in F—that is, functions whose "distance" from the target (as measured by R) is almost optimal. Then, one has to consider two possibilities: if the diameter of that set is small, the empirical minimization algorithm will preform very well on its convex hull, giving us the fast error rate we hope for. On the other hand, if the diameter of that set is large, there will be a major gain in the approximation error by considering functions in the convex hull. We will show that the set $\hat{F}_1$ is an empirical version of the set we would have liked to have.

Lemma 4.4 *There exists a constant c depending only on $\|\ell\|_{\mathrm{lip}}$ and b for which the following holds. Let F, x, α and $\hat{F}_1$ be defined as above. Then, with ν^n-probability at least $1 - 2\exp(-x)$, the best element f^F in the class F belongs to $\hat{F}_1$ and any function f in $\hat{F}_1$ satisfies*

$$R(f) \leq R(f^F) + c \max\left\{\alpha d(\hat{F}_1), (1 + b)\alpha^2\right\},$$

where $d(\hat{F}_1) = \mathrm{diam}(\hat{F}_1, L_2(\mu))$.

Proof Let $\mathcal{L}_f$ be the excess loss function associated with f (relative to F) and recall that f^F minimizes $h \longmapsto \mathbb{E}\ell(h(X), Y)$ in F. By the second part of Lemma 3.4, with μ^n-probability at least $1 - \exp(-x)$, every $f, g \in F$ satisfy

$$\left|\|f - g\|_{L_2^n}^2 - \|f - g\|_{L_2(\mu)}^2\right| \leq c_1 \max\left\{\|f - g\|_{L_2(\mu)} b\alpha, b^2\alpha^2\right\},$$

and hence, with that probability,

$$\|f - g\|_{L_2(\mu)}^2 \leq c_2 \max\left\{\alpha^2 b^2, \|f - g\|_{L_2^n}^2\right\}, \tag{4.3}$$

and

$$\|f-g\|_{L_2^n}^2 \leq c_3 \max\left\{\alpha^2 b^2, \|f-g\|_{L_2(\mu)}^2\right\}.$$

Also, by the first part of Lemma 3.4, with ν^n-probability at least $1-\exp(-x)$, for every $f \in F$

$$\left|\frac{1}{n}\sum_{i=1}^n \mathcal{L}_f(X_i,Y_i) - \mathbb{E}\mathcal{L}_f(X,Y)\right| \leq c_4\|\ell\|_{\mathrm{lip}} \max\left\{d_f\alpha, b\alpha^2\right\}, \tag{4.4}$$

where $d_f^2 = \|f-f^F\|_{L_2(\mu)}^2$.

Let $\mathcal{A}$ be the event in $(\Omega\times\mathbb{R})^n$ on which both (4.3) and (4.4) hold, and clearly $\nu^n(\mathcal{A}) \geq 1-2\exp(-x)$. Using (4.3), it is evident that on $\mathcal{A}$

$$d_f^2 \leq c_2 \max\left\{\alpha^2 b^2, \|f-f^F\|_{L_2^n}^2\right\},$$

and thus, by (4.4), for every $f \in F$

$$\left|\frac{1}{n}\sum_{i=1}^n \mathcal{L}_f(X_i,Y_i) - \mathbb{E}\mathcal{L}_f(X,Y)\right| \leq c_5\|\ell\|_{\mathrm{lip}} \max\left\{\alpha\|f-f^F\|_{L_2^n}, \alpha^2 b\right\}.$$

Moreover, since $\mathbb{E}\mathcal{L}_f \geq 0$, then on $\mathcal{A}$

$$\begin{aligned} P_n\ell_{f^F} &= P_n\ell_{\hat{f}} - P_n\mathcal{L}_{\hat{f}} \\ &\leq P_n\ell_{\hat{f}} - \mathbb{E}\left(\mathcal{L}_{\hat{f}}(X,Y)|\mathcal{D}\right) + \left|P_n(\mathcal{L}_{\hat{f}}) - \mathbb{E}\left(\mathcal{L}_{\hat{f}}(X,Y)|\mathcal{D}\right)\right| \\ &\leq P_n\ell_{\hat{f}} + c_5\|\ell\|_{\mathrm{lip}} \max\left\{\alpha\|\hat{f}-f^F\|_{L_2^n}, \alpha^2 b\right\}, \end{aligned}$$

where $\mathcal{D} = (X_i,Y_i)_{i=1}^n$. Therefore, if one chooses C_1 properly, f^F belongs to $\hat{F}_1$ on $\mathcal{A}$, i.e., with ν^n-probability greater than $1-2\exp(-x)$.

Next, let $d = \mathrm{diam}(\hat{F}_1, L_2(\mu))$. By the first part, $f^F \in \hat{F}_1$ on $\mathcal{A}$, and hence, on that event, $d_f \leq d$ for every $f \in \hat{F}_1$. Note that for every $f \in F$ and any sample $(X_i,Y_i)_{i=1}^n$,

$$\begin{aligned} R(f) &= R(f^F) + (P-P_n)(\mathcal{L}_f) + (P_n\ell_f - P_n\ell_{f^F}) \\ &\leq R(f^F) + (P-P_n)(\mathcal{L}_f) + (R_n(f) - R_n(\hat{f})). \end{aligned}$$

Thus, by the definition of $\hat{F}_1$ and the uniform estimates on $|(P_n-P)(\mathcal{L}_f)|$ in (4.4), it is evident that with ν^n-probability greater than $1-2\exp(-x)$,

$$R(f) \leq R(f^F) + c_4\|\ell\|_{\mathrm{lip}} \max\left\{d_f\alpha, b\alpha^2\right\} + C_1 \max\left\{\|\hat{f}-f\|_{L_2^n}\alpha, \alpha^2\right\},$$

for every $f \in \hat{F}_1$.

To complete the proof observe that since $\hat{f} \in \hat{F}_1$ then

$$\|\hat{f} - f\|_{L_2^n}^2 \leq c_3 \max\left\{\alpha^2 b^2, \|\hat{f} - f\|_{L_2(\mu)}^2\right\} \leq c_3\left\{\alpha^2 b^2, d^2\right\}.$$

□

Now we may turn to the second part of the algorithm—empirical minimization with respect to D_2 on the convex hull of $\hat{F}_1$ (which is, of course, independent of D_2).

Proof of Theorem 4.2. Fix $x > 0$ and let $\hat{\mathcal{C}}_1$ denote the convex hull $\operatorname{conv} \hat{F}_1$. By Lemma 4.4, we may assume that $f^F \in \hat{F}_1$ and set $d = \operatorname{diam}(\hat{F}_1, L_2(\mu))$. Since $f^F \in \hat{F}_1$ then

$$\max_{f \in \hat{F}_1} \|f - f^F\|_{L_2(\mu)} \leq d \leq 2 \max_{f \in \hat{F}_1} \|f - f^F\|_{L_2(\mu)}, \tag{4.5}$$

and let f_1 be a function in $\hat{F}_1$ that maximizes $f \longmapsto \|f - f^F\|_{L_2(\mu)}$ in $\hat{F}_1$.

Consider the second half of the sample $D_2 = (X_i, Y_i)_{i=n+1}^{2n}$. On one hand, by Corollary 3.3, with probability at least $1 - \exp(-x)$ (relative to D_2), for every $v \in \hat{\mathcal{C}}_1$

$$\left|\frac{1}{n}\sum_{i=1+n}^{2n} \mathcal{L}_{\hat{\mathcal{C}}_1}(v)(X_i, Y_i) - \mathbb{E}\left(\mathcal{L}_{\hat{\mathcal{C}}_1}(v)(X, Y)|D_1\right)\right| \leq c_1 \|\ell\|_{\text{lip}} \max\left\{d\alpha, b\alpha^2\right\},$$

where $\mathcal{L}_{\hat{\mathcal{C}}_1}(v)(X, Y) = \ell(v(X), Y) - \ell(f^{\hat{\mathcal{C}}_1}(X), Y)$ is the excess loss function relative to $\hat{\mathcal{C}}_1$. Since $\tilde{f}$ minimizes the empirical risk in $\hat{\mathcal{C}}_1$ on D_2 then $\frac{1}{n}\sum_{i=n+1}^{2n} \mathcal{L}_{\hat{\mathcal{C}}_1}(\tilde{f})(X_i, Y_i) \leq 0$. Therefore,

$$\begin{aligned}
R(\tilde{f}) &\leq R(f^{\hat{\mathcal{C}}_1}) + \mathbb{E}\left(\mathcal{L}_{\hat{\mathcal{C}}_1}(\tilde{f})|D_1\right) - \frac{1}{n}\sum_{i=n+1}^{2n} \mathcal{L}_{\hat{\mathcal{C}}_1}(\tilde{f})(X_i, Y_i) \\
&\leq R(f^{\hat{\mathcal{C}}_1}) + c_1 \|\ell\|_{\text{lip}} \max\left\{d\alpha, b\alpha^2\right\} \\
&= R(f^F) + \left(c_1 \|\ell\|_{\text{lip}} \max\left\{d\alpha, b\alpha^2\right\} - \left(R(f^F) - R(f^{\hat{\mathcal{C}}_1})\right)\right) \\
&\equiv R(f^F) + \beta,
\end{aligned} \tag{4.6}$$

and it remains to show that $\beta \leq c(x)\frac{\log M}{n}$.

To that end, we shall bound $R(f^F) - R(f^{\hat{\mathcal{C}}_1})$ using the convexity properties of ℓ (Assumption 4.1). Indeed, recall that $f^F \in \hat{F}_1$ (with high probability with respect to D_1) and that $f_1 \in \hat{F}_1$ maximizes the $L_2(\mu)$ distance to f^F in $\hat{F}_1$. Consider the mid-point $f_2 \equiv (f_1 + f^F)/2 \in \hat{\mathcal{C}}_1$. By our convexity assumption on the loss, all functions u and v in $L_2(\nu)$ satisfy

$$\mathbb{E}_\nu \phi\left(\frac{u+v}{2}\right) \leq \frac{1}{2}\mathbb{E}_\nu \phi(u) + \frac{1}{2}\mathbb{E}_\nu \phi(v) - \delta_\phi(\|u - v\|_{L_2(\nu)}),$$

(where for every $h \in L_2(\nu)$, $\mathbb{E}_\nu h = \mathbb{E}h(X,Y)$). In particular, for $u(X,Y) = f^F(X) - Y$ and $v(X,Y) = f_1(X) - Y$, the mid-point is $(u(X,Y) + v(X,Y))/2 = f_2(X) - Y$. Hence, using the assumption on δ_ϕ,

$$\begin{aligned} R(f_2) &= \mathbb{E}\ell(f_2(X),Y) = \mathbb{E}\phi\left(\frac{f_1(X)+f^F(X)}{2} - Y\right) \\ &\le \frac{1}{2}\mathbb{E}\phi(f_1(X)-Y) + \frac{1}{2}\mathbb{E}\phi\left(f^F(X)-Y\right) - \delta_\phi\left(\|f_1 - f^F\|_{L_2(\mu)}\right) \\ &\le \frac{1}{2}R(f^F) + \frac{1}{2}R(f_1) - c_\phi\frac{d^2}{4}, \end{aligned}$$

where the expectations are taken conditioned on D_1. By Lemma 4.4, the function $f_1 \in \hat{F}_1$ satisfies

$$R(f_1) \le R(f^F) + c_2 \max\left\{\alpha d, (1+b)\alpha^2\right\},$$

and thus,

$$R(f^{\hat{C}_1}) \le R(f_2) \le R(f^F) + c_3 \max\left\{\alpha d, (1+b)\alpha^2\right\} - c_4 d^2.$$

Therefore,

$$\begin{aligned} \beta &= c_1\|\ell\|_{\rm lip}\max\left\{d\alpha, b\alpha^2\right\} - \left(R(f^F) - R(f^{\hat{C}_1})\right) \\ &\le c_5\|\ell\|_{\rm lip}\max\left\{\alpha d, (1+b)\alpha^2\right\} - c_4 d^2. \end{aligned}$$

Finally, if $d \ge (c_6\|\ell\|_{\rm lip} + b)\alpha$ then $\beta \le 0$, otherwise $\beta \le c_7(\|\ell\|_{\rm lip} + b)\alpha^2$. □

Remark 4.5 Although we presented our results when both the functions in F and Y are uniformly bounded, those may be extended to the unbounded case, assuming that a reasonable tail estimate is satisfied by the functions in F and by Y. One example of such a situation which was studied in [3], is a weighted ℓ_1 regularized method $\check{f}$ in the Gaussian framework; that is, $Y = g(X) + W$ for a centered Gaussian variable W, where g is the regression function of Y given X, and assuming that F and g are uniformly bounded.

If we denote the L_2^n metric by $\|\ \|_n$, then it was proved that

$$\mathbb{E}\|\check{f} - g\|_n^2 \le (1+\varepsilon)\min_{f\in F}\|f-g\|_n^2 + C(\varepsilon)\frac{\log(M\vee n)}{n}. \tag{4.7}$$

Since Y has a nice tail decay (gaussian) one can show that our results apply to this case as well (with a slightly different probability estimate and constants). The fact that we have been able to obtain an oracle inequality with the exact constant 1 instead of $1+\varepsilon$ allows us to obtain a similar result to (4.7) by replacing each $R(f)$ with the empirical version of it - but, of course, our $\tilde{f}$ is different from $\check{f}$. Since minimization over the

convex hull of F and ℓ_1-penalization are in a one-to-one correspondence, one may view our result as an improvement to the inequality from [3], and the likely reason for the improved result is that we minimize over the "correct" subset of the convex hull rather than the entire convex hull.

5 The lower bound

Here, we will present an example that shows that empirical minimization performed in the convex hull is very far from being an optimal aggregation method. For every integer n, we will construct a function class $F = F_n$ with $M = c_1\sqrt{n}$ functions, for which, with probability greater than $1 - \exp(-c_2\sqrt{n})$, the empirical minimizer $\hat{f}$ in $\mathcal{C} = \mathrm{conv}(F)$ satisfies

$$R(\hat{f}) \geq R(f^F) + \frac{c_3}{\sqrt{n}},$$

where c_1, c_2 and c_3 are absolute constants.

Let $\Omega = [0, 1]$ endowed with the Lebesgue measure μ and set L_2 to be the corresponding L_2 space. Let $(\phi_i)_{i=1}^{\infty}$ be a realization of independent, symmetric, $\{-1, 1\}$-valued random variables as functions on $[0, 1]$ (for example, $(\phi_i)_{i=1}^{\infty}$ are the Rademacher functions). In particular, $(\phi_i)_{i=1}^{\infty}$ is an orthonormal family in L_2 consisting of functions bounded by 1. Moreover, the functions $(\phi_i)_{i=1}^{\infty}$ are independent and have mean zero.

Let M be an integer to be specified later and put $\ell(x, y) = (x - y)^2$. Consider

$$F = \{0, \pm\phi_1, \ldots, \pm\phi_M\}$$

and let $Y = \phi_{M+1}$ which is a noiseless target function. A sample is $(X_i, Y_i)_{i=1}^n = (X_i, \phi_{M+1}(X_i))_{i=1}^n$ where the X_i's are selected independently according to μ. It is clear that

$$\mathcal{C} = \mathrm{conv}(F) = \left\{ \sum_{j=1}^{M} \lambda_j \phi_j, \quad \sum_{j=1}^{M} |\lambda_j| \leq 1 \right\},$$

and that the true minimizers $f^F = f^{\mathcal{C}} = 0$; in particular, $R(f^F) = R(f^{\mathcal{C}})$ and there is no gain in the approximation error by considering functions in the convex hull $\mathcal{C}$. Also, the excess loss function of a function f, relative to F and to $\mathcal{C}$, satisfies

$$\mathcal{L}_f = (f - \phi_{M+1})^2 - (0 - \phi_{M+1})^2 = f^2 - 2f\phi_{M+1}.$$

Let $\Phi(x) = (\phi_i(x))_{i=1}^M$ and set $\langle \cdot, \cdot \rangle$ to be the standard inner product in $\ell_2^M = (\mathbb{R}^M, \|\cdot\|)$. Observe that Φ is a vector with independent $\{-1, 1\}$ entries, and thus, for every $\lambda \in \mathbb{R}^M$, $\mathbb{E}\langle \lambda, \Phi \rangle^2 = \|\lambda\|^2$; moreover, if we set $f_\lambda = \langle \lambda, \Phi \rangle$, then f_λ and ϕ_{M+1} are independent and since $\mathbb{E}\phi_{M+1} = 0$, the excess risk of f_λ satisfies

$$\mathbb{E}\mathcal{L}_{f_\lambda} = \mathbb{E}f_\lambda^2 = \|\lambda\|^2.$$

A significant part of our analysis is based on concentration properties of sums of random variables that belong to an Orlicz space.

Definition 5.1 For any $\alpha \geq 1$ and any random variable f, the ψ_α norm of f is

$$\|f\|_{\psi_\alpha} = \inf\{C > 0 : \mathbb{E}\exp(|f|^\alpha / C^\alpha) \leq 2\}.$$

The ψ_α norm measures the tail behavior of a random variable. Indeed, one can show that for every $u \geq 1$,

$$Pr(|f| > u) \leq 2\exp(-cu^\alpha / \|f\|_{\psi_\alpha}^\alpha),$$

where c is an absolute constant, independent of f [24].

The following lemma is a ψ_1 version of Bernstein's inequality [24].

Lemma 5.2 *Let $Y, Y_1, \ldots, Y_n$ be i.i.d random variables with $\|Y\|_{\psi_1} < \infty$. Then, for any $u > 0$,*

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^n Y_i - \mathbb{E}Y\right| > u\|Y\|_{\psi_1}\right) \leq 2\exp\left(-C_3 n \min\left(u^2, u\right)\right), \tag{5.1}$$

where $C_3 > 0$ is an absolute constant.

In the next lemma, we will present simple ψ_1 estimates for f^2 and the resulting deviation inequalities using Lemma 5.2.

Lemma 5.3 *There is an absolute constant C_4 for which the following holds. For every $\lambda \in \mathbb{R}^M$, $\|f_\lambda^2\|_{\psi_1} \leq C_4\|\lambda\|^2$ and for every $u > 0$,*

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^n f_\lambda^2(X_i) - \mathbb{E}f_\lambda^2\right| \geq uC_4\|\lambda\|^2\right) \leq 2\exp\left(-C_3 n \min\{u^2, u\}\right). \tag{5.2}$$

Proof Fix $\lambda \in \mathbb{R}^M$. Using Höffding's inequality [24] and the fact that $(\phi_i)_{i=1}^M$ are independent and symmetric Bernoulli variables, it follows that for every $u > 0$,

$$Pr\left(\left|\sum_{j=1}^M \lambda_j \phi_j\right| > u\|\lambda\|\right) \leq 2\exp(-u^2/2).$$

Hence, $\|\langle\lambda, \Phi\rangle\|_{\psi_2} \leq c_1\|\lambda\|$ for some absolute constant c_1. The first part of the lemma is evident because

$$\|f_\lambda^2\|_{\psi_1} = \|\langle\lambda, \Phi\rangle^2\|_{\psi_1} = \|\langle\lambda, \Phi\rangle\|_{\psi_2}^2 \leq c_1^2\|\lambda\|^2.$$

The second part of the claim follows from the first one and Lemma 5.2. □

Lemma 5.3 allows us to control the deviation of the empirical L_2^n norm from the actual L_2 norm for a large number of functions in a subset of $\{f_\lambda : \lambda \in S^{M-1}\}$. The subset we will be interested in is a maximal ε-separated subset of S^{M-1} for the right choice of $\varepsilon < 1$.

Lemma 5.4 *There exist absolute constants C_5, C_6, C_7 and C_8 for which the following hold. For any $n \geq C_5 M$, with μ^n-probability at least $1-2\exp(-C_6 n)$, for any $\lambda \in \mathbb{R}^M$,*

$$\frac{1}{2}\|\lambda\|^2 \leq \frac{1}{n}\sum_{i=1}^n f_\lambda^2(X_i) \leq \frac{3}{2}\|\lambda\|^2.$$

Also, for every $r > 0$, with μ^n-probability at least $1 - 6\exp(-C_6 M)$,

$$C_7\sqrt{\frac{rM}{n}} \leq \sup_{\{\lambda:\|\lambda\|\leq\sqrt{r}\}} \frac{1}{n}\sum_{i=1}^n f_\lambda(X_i)\phi_{M+1}(X_i) \leq C_8\sqrt{\frac{rM}{n}}.$$

Proof The proof of the first part is standard and we will sketch it for the sake of completeness. Since $f_\lambda = \langle \lambda, \Phi\rangle$, what we wish to prove is that with high probability,

$$\sup_{\lambda\in S^{M-1}} \left|\frac{1}{n}\sum_{i=1}^n \langle \lambda, \Phi(X_i)\rangle^2 - 1\right| \leq \frac{1}{2},$$

where S^{M-1} is the unit sphere in ℓ_2^M.

By a successive approximation argument [21], it is enough to prove that any point x in a maximal ε-separated subset N_ε of S^{M-1} (for an appropriate choice of ε), satisfies

$$\left|\frac{1}{n}\sum_{i=1}^n \langle x, \Phi(X_i)\rangle^2 - 1\right| \leq \delta$$

where ε and δ depend only on the constant $1/2$.

A volumetric estimate [21] shows that the cardinality of N_ε is at most $(5/\varepsilon)^M$. Hence, if we take $u = \delta/C_4$ in (5.2), then

$$Pr\left(\exists x \in N_\varepsilon : \left|\frac{1}{n}\sum_{i=1}^n \langle x, \Phi(X_i)\rangle^2 - 1\right| \geq \delta\right) \leq \left(\frac{5}{\varepsilon}\right)^M \cdot 2\exp\left(-C_3 n\delta^2/C_4^2\right)$$
$$\leq 2\exp(-c_0 n)$$

as long as $n \geq c_1(\varepsilon, \delta)M$.

Turning to the second part, since Φ is a vector of independent, symmetric Bernoulli variables and ϕ_{M+1} is also a symmetric Bernoulli variable, independent of the others, the supremum $\sup_{\{\lambda:\|\lambda\|\leq\sqrt{r}\}} \sum_{i=1}^n f_\lambda(X_i)\phi_{M+1}(X_i)$ has the same distribution as

$$\sup_{\{\lambda:\|\lambda\|\leq\sqrt{r}\}} \sum_{i=1}^n \varepsilon_i \langle \lambda, W_i\rangle = (*),$$

where $(\varepsilon_i)_{i=1}^n$ are symmetric Bernoulli variables that are independent of $(W_i)_{i=1}^n$, which are independent uniform random vertices of $\{-1,1\}^M$. Indeed, we can set $W_i := \Phi(X_i)$ and $\varepsilon_i := \phi_{M+1}(X_i)$. Clearly, for every $1 \le i \le n$, $\|W_i\|^2 = M$, and by the Kahane–Khintchine inequality [14]

$$\mathbb{E}_\varepsilon(*) = \sqrt{r}\mathbb{E}_\varepsilon \Big\| \sum_{i=1}^n \varepsilon_i W_i \Big\| \ge c_2\sqrt{r}\left(\sum_{i=1}^n \|W_i\|^2\right)^{1/2} = c_2\sqrt{rnM}.$$

Also,

$$\mathbb{E}_\varepsilon(*) \le \left(\mathbb{E}_\varepsilon(*)^2\right)^{1/2} \le \sqrt{rnM}.$$

To obtain the high probability estimate, we use the concentration result for vector valued Rademacher processes [14, Chapter 4]. Consider the ℓ_2^M-valued variables $Z = \sum_{i=1}^n \varepsilon_i \Phi(X_i)$ and let Z' be Z conditioned on $X_1, \dots, X_n$. By the first part of our claim, for any $n \ge c_1 M$ there is a set $\mathcal{A}$ with probability at least $1 - 2\exp(-c_0 n)$ on which $\sum_{i=1}^n \langle \lambda, \Phi(X_i)\rangle^2 \le (3/2)n\|\lambda\|^2$ for every $\lambda \in \mathbb{R}^M$. Thus, on $\mathcal{A}$,

$$\sigma^2(Z') \equiv \sup_{\theta \in S^{M-1}} \mathbb{E}_\varepsilon \left\langle Z', \theta \right\rangle^2 = \sup_{\theta \in S^{M-1}} \sum_{i=1}^n \langle \theta, \Phi(X_i)\rangle^2 \le \frac{3}{2}n,$$

implying that for any $u > 0$,

$$Pr\left(\left|\|Z'\| - \mathbb{E}_\varepsilon\|Z'\|\right| \ge u\sqrt{n}\right) \le 4\exp\left(-c_3 u^2\right),$$

where c_3 is an absolute constant. Since $n \ge c_1 M$ and $\mathbb{E}_\varepsilon(*) = \sqrt{r}\mathbb{E}_\varepsilon\|Z'\|$, it follows that if one takes $u = c_2\sqrt{M}/2$, there is an absolute constant c_4 for which, with probability at least $1 - 4\exp(-c_4 M)$,

$$\frac{c_2}{2}\sqrt{\frac{rM}{n}} \le \sup_{\{\lambda : \|\lambda\| \le \sqrt{r}\}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle \lambda, \Phi(X_i)\rangle \le 2\sqrt{\frac{rM}{n}}.$$

Therefore, combining the two high probability estimates, it is evident that with probability greater than $1 - 6\exp(-c_5 M)$,

$$\frac{c_2}{2}\sqrt{\frac{rM}{n}} \le \sup_{\{\lambda : \|\lambda\| \le \sqrt{r}\}} \frac{1}{n}\sum_{i=1}^n (f_\lambda \phi_{M+1})(X_i) \le 2\sqrt{\frac{rM}{n}}.$$

□

Now, we can formulate and prove the main result of this section, which will complete the proof of Theorem A.

Theorem 5.5 *There exist absolute constants c_1, c_2 and c_3 for which the following holds. For every integer n, set M such that $n = c_1 M^2$ and put F and $\mathcal{C}$ as defined above. Then, with μ^n-probability at least $1 - 8\exp(-c_2\sqrt{n})$, the empirical minimizer $\hat{f}$ in $\mathcal{C}$ satisfies that*

$$\mathbb{E}\mathcal{L}_{\hat{f}} \geq \frac{c_3}{\sqrt{n}}.$$

In particular, for that choice of M and n and with that probability, empirical minimization performed in $\mathcal{C}$ satisfies

$$R(\hat{f}) \geq \min_{f \in F} R(f) + \frac{c_3}{\sqrt{n}}.$$

Proof Fix $f_\lambda = \sum_{j=1}^M \lambda_j \phi_j \in \mathcal{C} = \mathrm{conv}(F)$ and recall that S^{M-1} is the unit sphere in ℓ_2^M. Note that $\mathbb{E}\mathcal{L}_{f_\lambda} = \sum_{i=1}^M \lambda_i^2$ and thus

$$\mathcal{L}_r \equiv \left\{\mathcal{L}_f : f \in \mathcal{C}, \mathbb{E}\mathcal{L}_f = r\right\} = \left\{\mathcal{L}_{f_\lambda} : \lambda \in B_1^M \cap \sqrt{r} S^{M-1}\right\} = \left\{\mathcal{L}_{f_\lambda} : \lambda \in \sqrt{r} S^{M-1}\right\},$$

provided that $r \leq 1/M$.

Since $\mathcal{L}_f = f^2 - 2f\phi_{M+1}$ then for every $r \leq 1/M$,

$$\begin{aligned}
&\inf_{\mathcal{L}_f \in \mathcal{L}_r} P_n \mathcal{L}_f \\
&= r - \sup_{\lambda \in \sqrt{r} S^{M-1}} (P - P_n)\mathcal{L}_{f_\lambda} \\
&= r - \sup_{\lambda \in \sqrt{r} S^{M-1}} \left(\left(\mathbb{E} f_\lambda^2 - \frac{1}{n}\sum_{i=1}^n f_\lambda^2(X_i) \right) - \frac{1}{n}\sum_{i=1}^n (-2 f_\lambda \Phi_{M+1})(X_i) \right) \\
&\leq r + \sup_{\lambda \in \sqrt{r} S^{M-1}} \left| \mathbb{E} f_\lambda^2 - \frac{1}{n}\sum_{i=1}^n f_\lambda^2(X_i) \right| - 2 \sup_{\lambda \in \sqrt{r} S^{M-1}} \frac{1}{n}\sum_{i=1}^n (f_\lambda \Phi_{M+1})(X_i).
\end{aligned}$$

Fix $r \leq 1/M$ to be named later. Applying both parts of Lemma 5.4, it follows that if $n \geq C_5 M$ then with probability at least $1 - 8\exp(-C_6 M)$,

$$\inf_{\mathcal{L}_f \in \mathcal{L}_r} P_n \mathcal{L}_f \leq \frac{3}{2} r - 2C_7 \sqrt{\frac{rM}{n}} = \sqrt{r}\left(\frac{3}{2}\sqrt{r} - 2C_7 \sqrt{\frac{M}{n}} \right).$$

Let $n = c_1 M^2$ and note that the condition that $n \geq C_5 M$ is satisfied. Hence, $\sqrt{M/n} = 1/\sqrt{c_1 M}$ and there are absolute constants $c_2 < 1$ and c_3, such that for $r \leq c_2/M$, $\inf_{\mathcal{L}_f \in \mathcal{L}_r} P_n \mathcal{L}_f \leq -c_3\sqrt{r/M}$.

On the other hand, combining the upper bounds from Lemma 5.4, it follows that for every $0 < \rho \leq 1/M$, with probability at least $1 - 8\exp(-C_6 M)$,

$$\sup_{\{\lambda:\|\lambda\|\le\sqrt{\rho}\}}\left|\frac{1}{n}\sum_{i=1}^n \mathcal{L}_{f_\lambda}(X_i)-\mathbb{E}\mathcal{L}_{f_\lambda}\right|$$
$$\le \sup_{\{\lambda:\|\lambda\|\le\sqrt{\rho}\}}\left|\frac{1}{n}\sum_{i=1}^n f_\lambda^2(X_i)-\mathbb{E}f_\lambda^2\right|+2\sup_{\{\lambda:\|\lambda\|\le\sqrt{\rho}\}}\left|\frac{1}{n}\sum_{i=1}^n (f_\lambda\phi_{M+1})(X_i)\right|$$
$$\le \frac{\rho}{2}+2C_8\sqrt{\frac{\rho M}{n}}=\frac{\rho}{2}+c_4\sqrt{\frac{\rho}{M}}\le c_5\sqrt{\frac{\rho}{M}}.$$

Therefore, on that set,

$$\inf_{\{\lambda:\|\lambda\|\le\sqrt{\rho}\}} P_n\mathcal{L}_{f_\lambda}\ge -\sup_{\{\lambda:\|\lambda\|\le\sqrt{\rho}\}}|(P_n-P)(\mathcal{L}_{f_\lambda})|\ge -c_5\sqrt{\frac{\rho}{M}}.$$

Hence, with probability at least $1-8\exp(-C_6M)$, if $c_5^2\rho\le c_3^2 r/2$, $\operatorname{argmin}_{f\in C}P_n\mathcal{L}_f$ is a function f_λ indexed by λ of norm larger than $\sqrt{\rho}$. In particular, such a function will have an excess risk greater than ρ. Therefore, taking $\rho\sim r$ and noting that one may select $r\sim 1/\sqrt{n}$, there exists an absolute constant $c_6>0$ such that with high probability,

$$\mathbb{E}\left(\mathcal{L}_{\hat{f}}|(X_i)_{i=1}^n\right)\ge\rho\ge\frac{c_6}{\sqrt{n}},$$

as claimed. □

References

1. Bartlett, P.L., Jordan, M.I., McAuliffe, J.D.: Convexity, classification, and risk bounds. J. Am. Stat. Assoc. **101**(473), 138–156 (2006)
2. Bunea, F., Nobel, A.: Sequential procedures for aggregating arbitrary estimators of a conditional mean. IEEE Trans. Inf. Theory **54**(4), 1725–1735 (2008)
3. Bunea, F., Tsybakov, A.B., Wegkamp, M.H.: Aggregation for Gaussian regression. Ann. Statist. **35**(4), 1674–1697 (2007)
4. Catoni, O.: Statistical learning theory and stochastic optimization, vol. 1851 of Lecture Notes in Mathematics. Springer, Berlin, 2004. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001
5. Dalalyan, A., Tsybakov, A.: Aggregation by exponential weighting, sharp oracle inequalities and sparsity. Mach. Learn. **72**(1–2), 39–61 (2008)
6. Dudley, R.M.: Uniform central limit theorems. Cambridge Studies in Advanced Mathematics, vol 3. Cambridge University Press, Cambridge (1999)
7. Gaïffas, S., Lecué, G.: Optimal rates and adaptation in the single-index model using aggregation. Electron. J. Stat. **1**, 538–573 (2007)
8. Giné, E., Zinn, J.: Some limit theorems for empirical processes. Ann. Probab. **12**(4), 929–998 (1984)
9. Guédon, O., Mendelson, S., Pajor, A., Tomczak-Jaegermann, N.: Subspaces and orthogonal decompositions generated by bounded orthogonal systems. Positivity **11**(2), 269–283 (2007)
10. Juditsky, A.B., Rigollet, P., Tsybakov, A.B.: Learning by mirror averaging. Ann. Statist. Available at http://www.imstat.org/aos/future_papers.html (2006, to appear)
11. Koltchinskii, V.: Local rademacher complexities and Oracle inequalities in risk minimization. Ann. Statist. **34**(6), 1–50, December 2006. 2004 IMS Medallion Lecture

12. Lecué, G.: Suboptimality of penalized empirical risk minimization in classification. In: Proceedings of the 20th Annual Conference On Learning Theory, COLT07. Lecture Notes in Artificial Intelligence, **4539**, 142–156, 2007. Springer, Heidelberg
13. Ledoux, M.: The concentration of measure phenomenon. Mathematical Surveys and Monographs, vol 89. American Mathematical Society, Providence, RI, 2001
14. Ledoux, M., Talagrand, M.: Probability in Banach spaces, volume 23 of Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]. Springer, Berlin (1991)
15. Lee, W.S., Bartlett, P.L., Williamson, R.C.: The importance of convexity in learning with squared loss. IEEE Trans. Inf. Theory **44**(5), 1974–1980 (1998)
16. Lugosi, G., Wegkamp, M.: Complexity regularization via localized random penalties. Ann. Statist. **32**(4), 1679–1697 (2004)
17. Mendelson, S.: Lower bounds for the empirical minimization algorithm. IEEE Trans. Inf. Theory (2007, to appear)
18. Mendelson, S.: On weakly bounded empirical processes. Math. Ann. **340**(2), 293–314 (2008)
19. Mendelson, S., Pajor, A., Tomczak-Jaegermann, N.: Reconstruction and subgaussian operators in asymptotic geometric analysis. Geom. Funct. Anal. **17**(4), 1248–1282 (2007)
20. Nemirovski, A.: Topics in non-parametric statistics. In Lectures on probability theory and statistics (Saint-Flour, 1998), vol 1738 of Lecture Notes in Math., pages 85–277. Springer, Berlin, 2000
21. Pisier, G.: The volume of convex bodies and Banach space geometry. Cambridge Tracts in Mathematics, vol 94. Cambridge University Press, Cambridge (1989)
22. Tsybakov, A.B.: Optimal rates of aggregation. In: Proceedings of the 16th Annual Conference On Learning Theory, COLT03. Lecture Notes in Artificial Intelligence, **2777**, 303–313, 2003. Springer, Heidelberg
23. Tsybakov, A.B.: Introduction à l'estimation non-paramétrique. Springer, Berlin, 2004
24. van der Vaart, A.W., Wellner, J.A.: Weak convergence and empirical processes, Springer Series in Statistics. Springer, New York (1996)